

**SYSTEM AND METHOD FOR INTERLEAVING  
INFINIBAND SENDS AND RDMA READ  
5 RESPONSES IN A SINGLE RECEIVE QUEUE**

**Inventor:** James A. Mott & Elisa Rodrigues

**BACKGROUND**

10 This invention relates to the field of computer systems. More particularly, a system and methods are provided for accommodating interleaved Sends and RDMA Read responses in a single InfiniBand receive queue.

InfiniBand™ technology provides a flexible, scalable architecture for interconnecting servers, communication networks, storage components and other  
15 systems and devices. Computing and storage nodes have become distributed throughout many organizations' computing environments, and the InfiniBand architecture provides means for interconnecting those elements and others. For example, InfiniBand channel adapters can be used as bridges between an InfiniBand fabric and external communication systems or networks.

20 In the InfiniBand architecture, a queue pair (QP) defines an end-to-end connection between two nodes (e.g., servers, input/output components) at the transport protocol layer. A virtual lane (VL) operates at the link layer, and defines single-hop connections (e.g., between two switches, between a switch and a node). Each virtual lane has an associated service level indicating a quality of  
25 service to be afforded the traffic within that virtual lane. When an InfiniBand packet is communicated, it is communicated as part of a specific queue pair, which is assigned membership in a virtual lane for each hop. The virtual lanes

used for different hops may vary, but the different virtual lanes may be associated with the same service level.

Queue pairs are flow-controlled by the receiving end of the end-to-end connection. Virtual lanes are flow-controlled by the receiving end of each hop. In particular, a node that receives traffic via an end-to-end connection or single hop may issue credits allowing the transmitting end (of the connection or hop) to send a specified amount of traffic.

A QP credit is generally issued for each message (e.g., one credit equals one message of up to  $2^{32}$  bytes), and each message may be segmented into one or more InfiniBand packets. For example, one message may correspond to one Ethernet packet to be encapsulated in one or more InfiniBand packets and passed to an external network. VL credits are generally in the form of blocks (e.g., sixty-four bytes per credit). When the receiving end of a QP or VL issues a credit, it is generally understood that an amount of storage space sufficient to store the corresponding amount of traffic will be available when the traffic is received. If storage space is not available, the QP at the receiving end may instruct the sender to retry the communication later. A VL reports a flow control error and then drops the packet.

The InfiniBand specification implies that each QP and each VL should be serviced at its receiving end by a separate FIFO (First-In, First-Out) queue. However, providing dedicated queues requires each queue pair and virtual lane to be provided with worst-case buffering to accept a maximum burst of traffic. This scheme results in an inefficient use of memory space because, at any given time, not every active QP or VL will even be configured, much less receiving enough traffic to require a full set of buffers, and therefore storage space dedicated to a particular (e.g., non-busy) QP or VL may be wasted. Thus, a need exists for a

system and method for sharing buffers between multiple queue pairs or multiple virtual lanes, and/or between queue pairs and virtual lanes.

A shared storage space for virtual lane and queue pair traffic may allow more flexibility and scalability, but it would still be necessary to support flow control. For example, with shared storage space, the amount of storage used by each VL and QP should be tracked in order to calculate how many credits the receiving end can or should issue. Depending on whether any storage space is dedicated to a queue pair or virtual lane, or how much shared space is available for use by any queue pair or virtual lane, supporting flow control may become problematic. Thus, there is a need for a system and method for facilitating flow control in association with a memory configured for shared buffering of queue pairs and/or virtual lanes.

Further, at an interconnection between an InfiniBand fabric and an external system (e.g., an Ethernet network or other communication system), the use of discrete FIFO queues for each terminating QP (and/or VL) means that traffic to be transferred from a QP to the external system must be copied from its InfiniBand QP queue into a different queue or data structure for the external system (e.g., a network transmit module) before the traffic can be transmitted externally. This delays the transfer and causes additional inefficiency. Thus, there is a need for a system and method for avoiding inefficient memory operations when transferring communications between InfiniBand and an external system.

Also, if a single receive queue is used to store mixed types of traffic for a queue pair or other type of communication connection, a system and method are needed for interleaving the different types of traffic while avoiding the possibility of transferring traffic out of order. For example, a queue pair's traffic may include Send commands containing encapsulated outbound communications (e.g.,

Ethernet packets), Send commands containing RDMA Read descriptors (e.g., for retrieving outbound communications), responses to RDMA Reads, etc. Thus, different types of traffic should be handled without causing out of order processing of outbound communications.

- 5           A system and method are also needed to track responses to RDMA Read operations, so that a corresponding entry in a retry queue can be retired when all responses are received.

### SUMMARY

- 10           A system and method are provided for processing interleaved Sends and RDMA Read responses in a single InfiniBand queue pair receive queue. In one embodiment, the queue is implemented as one or more linked lists of memory buckets. The queue stores InfiniBand commands until their associated communications are assembled and forwarded to a transmit module. The
- 15    commands include Send commands containing encapsulated communications from a host, and/or Send commands encapsulating RDMA Read descriptors for retrieving and constructing a communication from a host.

- The queue grows as new InfiniBand packets and commands are received, and shrinks as the associated communications (e.g., Ethernet packets) are
- 20    forwarded to the transmit module. In one embodiment, the queue comprises two linked lists – one for receiving and queuing new Send commands, and one for reassembling the communications corresponding to the Send commands. Each linked list is managed by head and tail pointers. A next packet pointer identifies the next command whose communication should be processed (i.e., assembled
- 25    and forwarded).

          When a queued command is processed, if it is a Send command with an encapsulated communication, the communication is forwarded. Otherwise, for a

Send command encapsulating RDMA Read descriptors, RDMA Read requests corresponding to the descriptors are issued to the specified host. Responses to the Read requests bypass the reception area of the queue (e.g., a first linked list where newly received commands are enqueued) and are stored and assembled in an assembly area (e.g., a second linked list). For a Send command with encapsulated RDMA Read descriptors, the assembly area may be expanded as necessary to accommodate the communication.

### DESCRIPTION OF THE FIGURES

FIG. 1 is a block diagram depicting a computing device in which traffic received from multiple queue pairs and virtual lanes are buffered in a single shared memory, in accordance with an embodiment of the present invention.

FIG. 2 is a block diagram of memory structures for facilitating the combined buffering of queue pair and virtual lane traffic, in accordance with an embodiment of the invention.

FIG. 3 is a flowchart illustrating one method of storing InfiniBand traffic from multiple queue pairs and virtual lanes in a shared memory, in accordance with an embodiment of the present invention.

FIG. 4 is a block diagram of an InfiniBand receive module, according to one embodiment of the invention.

FIG. 5 is a block diagram of a flow control portion of an InfiniBand link core, showing its interaction with a Resource Manager, according to one embodiment of the invention.

FIG. 6 is a graph demonstrating one method of setting thresholds and corresponding amounts of advertisable message credits, according to one embodiment of the invention.

FIG. 7 is a flowchart illustrating one method of applying flow control to InfiniBand traffic received from multiple queue pairs and virtual lanes and stored in a shared memory structure, in accordance with an embodiment of the present invention.

5           FIG. 8 is a flowchart illustrating one method of mapping InfiniBand communications to an external communication system, in accordance with an embodiment of the present invention.

FIG. 9 is a block diagram of a queue pair queue configured to accommodate mixed types of traffic without causing out-of-order receipt,  
10           according to one embodiment of the invention.

FIG. 10 is a flowchart illustrating one method of processing traffic received in the queue pair queue of FIG. 9, according to one embodiment of the invention.

FIG. 11 is a block diagram of a memory structure for maintaining linked  
15           lists for tracking receipt of responses to RDMA Read operations, according to one embodiment of the invention.

FIG. 12 is a flowchart demonstrating one method of tracking receipt of responses to RDMA Read operations, according to one embodiment of the invention.

20

## DETAILED DESCRIPTION

The following description is presented to enable any person skilled in the art to make and use the invention, and is provided in the context of particular applications of the invention and their requirements. Various modifications to the disclosed embodiments will be readily apparent to those skilled in the art and the general principles defined herein may be applied to other embodiments and applications without departing from the scope of the present invention. Thus, the present invention is not intended to be limited to the embodiments shown, but is to be accorded the widest scope consistent with the principles and features disclosed herein.

The program environment in which a present embodiment of the invention is executed illustratively incorporates a general-purpose computer or a special purpose device such as a hand-held computer. Details of such devices (e.g., processor, memory, data storage, display) may be omitted for the sake of clarity.

It should also be understood that the techniques of the present invention may be implemented using a variety of technologies. For example, the methods described herein may be implemented in software executing on a computer system, or implemented in hardware utilizing either a combination of microprocessors or other specially designed application specific integrated circuits, programmable logic devices, or various combinations thereof. In particular, the methods described herein may be implemented by a series of computer-executable instructions residing on a suitable computer-readable medium. Suitable computer-readable media may include volatile (e.g., RAM) and/or non-volatile (e.g., ROM, disk) memory, carrier waves and transmission media (e.g., copper wire, coaxial cable, fiber optic media). Exemplary carrier waves may take the form of electrical, electromagnetic or optical signals

conveying digital data streams along a local network, a publicly accessible network such as the Internet or some other communication link.

### **Introduction**

5           In one embodiment of the invention, a system and method are provided for buffering traffic received via InfiniBand queue pairs (QP) and virtual lanes (VL) in a single shared memory structure. Memory buckets or elements are dynamically allocated as needed. For each active queue pair, a linked list of memory buckets is formed for storing traffic from that queue pair. Thus, instead  
10 of a fixed number of queues or fixed-size, dedicated queues (e.g., FIFO queues), each QP has a dynamically sized linked list that can be reconfigured to support varying numbers of virtual lanes and queue pairs, and each resulting linked list can be easily measured and manipulated. Each queue pair's virtual lane membership is noted, thereby facilitating measurement of the amount of traffic in  
15 the shared memory for each active virtual lane.

          In another embodiment of the invention, a system and method are provided for facilitating flow control of queue pairs and/or virtual lanes, wherein the queue pair and virtual lane traffic is buffered in a shared memory structure. In this embodiment, depths of the queue pairs and virtual lanes are measured, and a  
20 decision whether to accept a new packet, or issue a credit, may be made on the basis of whether there is sufficient room in the shared memory structure for the packet's queue pair or virtual lane.

          In yet another embodiment of the invention, a system and method are provided for sharing a memory between the receiving end of InfiniBand network communication connections (e.g., queue pairs, virtual lanes) and the transmitting  
25 end of a communication network system or link external to the InfiniBand network (e.g., an Ethernet network). The memory may be used for combined



buffering of receive queue pairs and/or virtual lanes via linked lists, but also comprises linked lists for one or more outbound ports. A communication can be queued for transmission (e.g., after being reassembled in the shared memory) by simply copying, moving or re-arranging pointer or register values, rather than  
5 copying the entire communication.

In another embodiment of the invention, a system and method are provided for mixing traffic having different transfer protocols in one queue, while avoiding out-of-order processing of the traffic. The queue may be for an InfiniBand queue pair, and may be implemented as one or more linked lists. For  
10 example, such a queue may store Send commands encapsulating outbound communications and Send commands encapsulating RDMA (Remote Direct Memory Access) Read descriptors for retrieving outbound communications. Each command is processed, and its associated communication (e.g., an Ethernet packet) forwarded for transmission, in order.

In another embodiment of the invention, a system and method are provided for tracking responses to an RDMA Read operation. Illustratively, linked lists may be maintained for different queue pairs, with each linked list entry storing the range of Packet Sequence Numbers (PSN) associated with the expected responses to the RDMA Read. When the last response is received, the  
15 linked list entry may be removed.

Embodiments of the invention are described below as they may be implemented for InfiniBand traffic traversing queue pairs and virtual lanes. Other embodiments of the invention may be configured and implemented for other types of communication architectures or protocols, such as PCI (Peripheral Component  
20 Interconnect) Express, Asynchronous Transfer Mode (ATM), Ethernet and, in general, any packetized data transfer scheme that multiplexes different

independent packet streams onto a shared medium using “Send” or RDMA protocols.

### **Combined Buffering of Virtual Lanes and Queue Pairs**

- 5           In one embodiment of the invention, a shared memory structure is used to store traffic received at a computing device from an InfiniBand fabric or network. The memory structure is shared among multiple queue pairs that define end-to-end connections between the computing device and other InfiniBand nodes. The queue pairs may be members of any virtual lane(s) reaching the computing device.
- 10   The shared memory structure is managed as a set of linked lists maintained outside the shared memory structure, thereby allowing traffic from multiple queue pairs and/or virtual lanes to be stored and reassembled in the same structure simultaneously. Thus, there is no need to maintain separate and/or dedicated memory structures for each queue pair and/or virtual lane. In one implementation,
- 15   the traffic is transmitted from the same structure after being reassembled into an outgoing communication (e.g., an Ethernet packet).

- In this embodiment, the amount of traffic stored for a given queue pair at a particular time may be measured by examining that queue pair’s linked list. A determination of the amount of traffic in the shared memory for one virtual lane
- 20   may be facilitated by monitoring the virtual lane membership of each queue pair and accumulating the sizes of the linked lists for each queue pair in the virtual lane.

- FIG. 1 is a block diagram of a computing device in which this and other embodiments of the invention may be implemented. Although these
- 25   embodiments are described as they may be configured for transferring communications between an InfiniBand fabric and an Ethernet network, other embodiments may be configured for interfacing between other types of

communication networks, systems or components, such as SONET (Synchronous Optical Network) for Internet Protocol (IP), POS (Packet over SONET), PCI Express or SONET/SDH (Synchronous Digital Hierarchy).

In FIG. 1, computing or communication device 102 is coupled to  
5 InfiniBand fabric 104, and is also coupled to Ethernet network 106 or some other external communication system or component. Any number of virtual lanes may be configured to be active between device 102 and switches or other link partners within fabric 104. Similarly, any number of queue pairs may be configured to be active between device 102 (or components of device 102) and other nodes within  
10 the InfiniBand fabric. The queue pairs may be reliable connected queue pairs (RCQPs).

Device 102 includes channel adapter 110 (e.g., a target channel adapter), which is configured for communication with nodes within InfiniBand fabric 104. Channel adapter 110 comprises control 112, memory 114, InfiniBand Link Core  
15 118, InfiniBand Receive Module (IRX) 120, Network Transmit Module (NTX) 130 and external port(s) 116. IRX 120 includes queue pair pointers 122 for queue pairs' linked lists, while NTX 130 includes transmit pointers 132 associated with queues for external ports 116. Through external port(s) 116, channel adapter 110 can transmit Ethernet packets and/or other communications onto network 106.

20 InfiniBand Link Core 118 performs link-level flow control using credit information, as described below.

IRX 120 handles incoming packets from InfiniBand fabric 104. Queue pair pointers 122 comprise registers or pointers for managing queues for each active queue pair. As described below, a queue pair's queue may comprise linked  
25 lists of entries in control 112 and memory 114. A queue pair's linked list is used to manage the reassembly, in memory 114, of segmented outbound communications from the contents of one or more InfiniBand packets processed

by IRX 120. IRX 120 also stores queue pair state information (possibly with queue pair pointers 122) and virtual lane state information. Further details of an InfiniBand Receive Module are described below, in conjunction with FIG. 4.

5 NTX 130 processes outbound communications after they are reassembled in memory 114, for transmission via an external port 116. An external port may offer Quality of Service (QoS) options by maintaining separate output queues for each defined QoS; access to the queues may be arbitrated via weighted fair queuing or other arbitration schemes.

10 Illustratively, transmit pointers 132 correspond to one or more linked lists of reassembled communications awaiting transmission. For example, for each queue (e.g., QoS queue) of an external port, transmit pointers 132 may include a separate set of registers or pointers for managing a linked list (within control 112 and memory 114) of communications awaiting transmission. Methods of transferring outbound communications between InfiniBand fabric 104 and  
15 network 106, through a channel adapter such as channel adapter 110, are discussed in a following section.

In one implementation of the embodiment of the invention depicted in FIG. 1, channel adapter 110 is capable of operating in any of two or more modes. In one mode, a single external port 116 is operated at a data rate of approximately  
20 10Gbps. In another mode, multiple (e.g., 4) external ports 116 are operated, each at a data rate of approximately 1Gbps.

In the illustrated embodiment of the invention, as InfiniBand packets are received at computing device 102, over various queue pairs and virtual lanes, contents of the packets are stored or reassembled via memory 114 and control  
25 112. Thus, one contiguous memory structure (i.e., memory 114) is used to buffer packets for multiple queue pairs and virtual lanes, rather than implementing a separate structure (e.g., FIFO queues) for each.

In this embodiment of the invention, control 112, memory 114, IRX 120 and NTX 130 reside on one chip or integrated circuit (e.g., an ASIC). In other embodiments, multiple chips may be used and/or the illustrated elements of the channel adapter may be configured differently.

5           FIG. 2 depicts details of a channel adapter such as channel adapter 110 of FIG. 1 in greater detail, according to one embodiment of the invention. Although the outbound path of communications (i.e., from InfiniBand to an external system) is discussed herein, the channel adapter is also configured to handle incoming communications (i.e., from the external system to the InfiniBand fabric).

10           In FIG. 2, memory 204 is a memory structure (e.g., a multi-port RAM) configured to store traffic received via various queue pairs and virtual lanes, from any number of InfiniBand nodes and link partners, for reassembly into communications destined for an external communication system or component. Each queue pair's traffic may be stored as a linked list of memory locations.

15           Control 202 is a separate memory or data structure for managing linked lists of each queue pair's traffic stored in memory 204.

          Illustratively, the traffic comprises contents of InfiniBand packets configured to carry encapsulated Ethernet packets from InfiniBand nodes for transmission on the external system. The packets may include InfiniBand Send

20           commands (encapsulating Ethernet packet segments or RDMA Read descriptors), RDMA Read commands and/or other types of packets.

          For each queue pair that is active on the channel adapter, queue pair pointers 210 include (at least) a head and tail pointer to identify the beginning and end of the queue pair's linked list. Thus, QP2 head 212a and QP2 tail 212b are

25           pointers for a linked list associated with a particular queue pair. Similarly, transmit pointers 220 includes a set of head and tail pointers for each queue of an outbound port. Queue pair pointers 210 and/or transmit pointers 220 may be

stored in registers or other structures. Other information may also be stored for each queue pair or outbound queue, such as the virtual lane that a queue pair belongs to, other pointers, etc.

Each queue pair's and outbound queue's linked list comprises a series of  
5 corresponding entries in control 202 and associated memory buckets in memory 204. A set of memory buckets in memory 204 corresponding to a particular linked list may be considered to constitute a queue (e.g., for the associated queue pair and/or virtual lane). A "bucket" in this context comprises a set of lines in a RAM that have a common value for their most significant address bits. The  
10 number of other, least significant, address bits may determine the size of a bucket (i.e., the number of lines or bytes in the bucket).

Control 202 and memory 204 may include an equal number of entries. Thus, in the implementation depicted in FIG. 2, control 202 comprises 1,024 (1K) entries, and memory 204 comprises 1,024 (1K) buckets. Each control entry  
15 includes a 10-bit value, which is used to identify the next control entry in the linked list and the corresponding next bucket. A control entry may also contain other information (e.g., an ECC code).

The control structure may be protected by a number of techniques, used in combination or separately. These techniques may include physical separation of  
20 bits in the same control entry in order to prevent double bit errors, implementation of a SECDED (Single Error Correct, Double Error Detect) Error Correction Code (ECC), etc. The SECDED code protection can be extended to include the address of the control structure entry, thereby protecting the data from being written or read from the wrong location due to an addressing error while accessing the  
25 control structure.

In FIG. 2, each data line in memory 204 is 128 bits (16 bytes) wide, and each bucket includes eight lines. Thus, the bucket size is 128 bytes, and the size of memory 204 is 128 Kbytes.

5 In other embodiments of the invention, control 202 and memory 204 may be of virtually any size and the configuration of a control entry and a bucket may vary. For example, the size of a bucket or a line in a bucket may be configured based on the size (e.g., average size) of an InfiniBand packet payload. In addition, the size and configuration of memory 204 may differ, depending on whether it is implemented on the same chip or component as other elements, or is implemented  
10 as an external memory. For example, an external memory of 4MB may be used, with bucket sizes of 2KB.

In FIG. 2, QP2 head 212a is the head pointer for a first queue pair, and therefore identifies the first control entry for QP2 – e.g., entry i. QP2 head 212a also identifies the first memory bucket for QP2 – bucket i. A value stored in  
15 control entry i (i.e., the value 1023) identifies the next control entry for the linked list and the next bucket – number 1023. Similarly, control entry 1023 identifies the next control entry and corresponding bucket (i.e., m). In the implementation of FIG. 2, control entry 0 is a null pointer. This entry may therefore be used to terminate a linked list.

20 QP2 tail 212b also identifies control entry m and bucket m, thus indicating where the next set of data (e.g., the payload of the next InfiniBand packet for QP2) should be stored. Information concerning how full a bucket is may be stored in queue pair pointers 210 or some other location.

For example, this information may be stored in the first line of the first  
25 bucket allocated to each outbound communication stored in the linked list. Additional information (e.g., queue pair number, virtual lane identifier, outbound port/queue identifier) may be stored before, after or as part of a communication

reassembled in memory 204. Such information may be used to help process or transmit the communication correctly.

Illustratively, InfiniBand payloads are stored contiguously, within and across buckets, from the beginning to the end of one outbound communication (e.g., one encapsulated Ethernet packet). When the end of one communication is reached, a new bucket (and new control memory entry) may be started for that queue pair's linked list. In an alternative embodiment of the invention, portions of multiple different outbound communications may be stored in a single memory bucket.

10 A given control entry may be considered to be associated with, or correspond to, the memory bucket at the same position in a queue pair's linked list. Thus, in FIG. 2, control entry i is associated with memory bucket i, control entry m corresponds to memory bucket m, and so on.

In an embodiment of the invention, linked lists of outbound communications (e.g., communications that have been reassembled), anchored by head and tail pointers in transmit pointers 220 (e.g., PortA head pointer 222a and PortA tail pointer 222b), are maintained in the same way. Thus, each of control 202 and memory 204 are multi-ported structures in FIG. 2.

Free head 230a and free tail 230b are pointers used to maintain a linked list of free control entries and buckets. Thus, free head 230a identifies a first free control entry and a corresponding first free bucket in a free list, while free tail 230b identifies the last control entry and the last free bucket in the list. Free control entries and buckets may be removed from the head (or first) end and returned to the tail (or last) end of the free list. Similarly, buckets may be removed from a queue pair's linked list of buckets at the head end and returned at the tail end.



FIG. 3 demonstrates a method of sharing a single memory structure to store InfiniBand traffic from multiple queue pairs and virtual lanes, according to one embodiment of the invention.

5 In operation 302, an InfiniBand packet is received at a channel adapter of a device coupled to an InfiniBand fabric and an external communication system (e.g., an Ethernet network). In this embodiment, the packet contains all or a portion of an encapsulated Ethernet packet.

In operation 304, InfiniBand packet validation rules are applied to determine whether the packet is valid.

10 In operation 306, if the packet is a valid InfiniBand packet the method continues to operation 308; otherwise, the method advances to operation 322.

In operation 308, the virtual lane and queue pair on which the InfiniBand packet was received are identified. Illustratively, the service level or virtual lane associated with the packet may affect the QoS afforded the Ethernet packet when it is transmitted on an outbound port.

15 In operation 310, the tail of the linked list for the packet's queue pair is located. In this embodiment, a pointer to the tail is maintained as part of a per-queue pair collection of data. The tail identifies the bucket in the shared memory in which the last packet payload was stored and the corresponding entry in a shared control structure that facilitates management of the linked list.

20 In operation 312, it is determined whether there is space in the shared memory structure for storing contents of the packet. If so, the illustrated method continues with operation 314; otherwise, the method advances to operation 322. A method of monitoring the amount of space in the memory for storing packets from different queue pairs is described in a following section.

In operation 314, the payload of the InfiniBand packet is stored in the memory structure. In this embodiment, entries in the shared control structure for

the shared memory structure contain information regarding the status of the bucket that corresponds to the entry. Thus, it can readily be determined where in the bucket the payload should be stored and whether the entire payload will fit in the bucket. Illustratively, if the payload comprises the first fragment or portion of an outbound communication, or a set of RDMA Read descriptors for retrieving a communication, a new bucket may be used. Otherwise, the payload is stored contiguously with the previous payload for the communication.

In operation 316, the queue pair's linked list and possibly other lists or information are updated as necessary. For example, if a new memory bucket is needed to accommodate the payload, a bucket may be taken from a list of free (available or empty) memory buckets. The bucket and a new corresponding entry in the control structure are initialized as necessary and added to the QP's linked list.

In operation 318, depth indicators (e.g., pointers, counters) for the packet's queue pair and/or virtual lane may be updated as necessary to indicate how much traffic is currently stored in the shared memory for the queue pair and/or virtual lane. Queue Pair and/or virtual lane credits may be issued if warranted. The illustrated method then ends.

In operation 322, the received packet is either invalid (e.g., contains an unrecoverable error) or there is no space to store the packet payload in the shared memory. In this embodiment of the invention, the packet is therefore discarded. In other embodiments, other action may be taken.

## **Receive Flow Control with Combined Buffering of Virtual Lanes and Queue Pairs**

In one embodiment of the invention, a system and method are provided for facilitating flow control of InfiniBand receive traffic, at the link layer and/or

transport layer, when traffic received via virtual lanes and queue pairs are buffered in a single memory structure. An implementation of this embodiment is suitable for use with shared buffering as described in the preceding section, and the channel adapter described in conjunction with FIGs. 1 and 2.

5           In this embodiment, to properly apply flow control to queue pairs and/or virtual lanes, the amount of traffic stored in the shared memory for each active queue pair and/or virtual lane is tracked. As described above, linked lists may be maintained for each queue pair, thereby facilitating each queue pair's usage of the shared memory. And, by noting each queue pair's virtual lane membership, the  
10          total memory usage of all queue pairs within a particular virtual lane can be easily calculated.

          In an embodiment of the invention, a dedicated portion of the shared memory structure may be allocated to a given queue pair or virtual lane. And, queue pairs and virtual lanes allocated a dedicated portion of the memory may or  
15          may not be permitted to also use a shared portion of the memory that can be drawn upon by multiple queue pairs and/or virtual lanes. Queue pairs and virtual lanes not allocated dedicated memory space use only shared memory for queuing their traffic. Thus, applying flow control for an individual queue pair or virtual lane will consider the amount of space available to it, including dedicated and/or  
20          shared space.

          In this embodiment, an InfiniBand Resource Manager (IRM) module, which may be part of an InfiniBand Receive Module (IRX), manages the shared memory. In particular, the IRM allocates or grants memory buckets (i.e., units of memory), receives buckets (e.g., after they are used), and performs other actions  
25          to facilitate usage of the shared memory structure. In particular, it tracks the amount of traffic in the shared memory for a channel adapter's queue pairs and virtual lanes. It also implements or facilitates various flow control mechanisms,

such as link layer (virtual lane) credits, transport layer (queue pair) credits, retries and RDMA Read operations. It also maintains various operating parameters regarding the queue pairs and virtual lanes.

FIG. 4 is a block diagram of an InfiniBand Receive Module (IRX),  
5 according to one embodiment of the invention. The IRX is part of a channel adapter or other device used to interface a computing or communication device to an InfiniBand fabric, and contains a shared memory structure providing common buffering for multiple queue pairs and/or virtual lanes.

In FIG. 4, IRX 402 includes InfiniBand Resource Module (IRM) 410,  
10 Receive Packet Processor (RPP) 412, Post Packet Processor (PPP) 414, acknowledgement generator 416, Link List (Receive) Manager (LLRM) 418 and CPU interface 420, in addition to a collection of queue pair pointers (not shown in FIG. 4).

IRM 410 includes queue pair memory (or memories) 430 and virtual lane  
15 memory (or memories) 432, and interfaces with Network Transmit Module (NTX) 404 and InfiniBand Link Core (ILC) 406. Not shown in FIG. 4 are the shared memory and shared control structures (e.g., memory 114 and control 112 of FIG. 1) in which queue pairs' queues (i.e., linked lists) are maintained.

In one embodiment of the invention, IRM 410 supports four virtual lanes  
20 for each receive port of the channel adapter on which the IRM resides, in addition to always-present virtual lane 15 (which is dedicated to management traffic). The channel adapter may operate in a single-port or dual-port mode.

In this embodiment, up to 64 user-assignable active queue pairs may be supported by IRM 410, with queue pairs zero and one being reserved for  
25 management traffic. In other embodiments, other quantities of virtual lanes and queue pairs may be supported by an InfiniBand Resource Module (e.g., up to  $2^{24}$ ).

In FIG. 4, RPP 412 requests and receives resources (e.g., memory buckets, entries in the control structure) from IRM 410 for new packets received at the channel adapter. For example, RPP 412 may notify IRM 410 of a need to store a payload of a new packet, and may indicate the amount of space needed. IRM 410  
5 may then allocate or reserve a sufficient number of memory buckets for the payload. The RPP also facilitates the storage of data received via RDMA Read operations, by queuing RDMA Read descriptors for example. RPP 412 may return unused buckets to the IRM if, for example, the packet is dropped or rejected (e.g., because of an error). RPP 412 may also recognize the late detection  
10 of an error in a packet and return allocated resources if the packet must be rejected.

PPP 414 evaluates the size of an RDMA Read descriptor queued by the RPP, and signals IRM 410 to reserve the necessary resources. The resources are reserved for that RDMA Read and, when the operation occurs, the RPP  
15 recognizes the operation and matches the reserved resources with the operation. Thus, an RNR-NAK should never need to be issued for an RDMA Read operation.

Acknowledgement generator 416 generates fields of an InfiniBand ACK (acknowledgment) or NAK (negative acknowledgment) packet, and constructs  
20 and forwards ACK packets of transport layer flow control information to an InfiniBand Transmit (ITX) module (e.g., to advertise message credits, issue an RNR-NAK). For example, the acknowledgement generator may query IRM 410 to determine how much storage space is available for a particular queue pair, and report a corresponding number of available queue pair credits to the ITX.

25 LLRM 418 maintains linked lists of used and unused memory buckets and control entries. Illustratively, IRM 410 passes returned buckets to LLRM 418 for

addition to the free list. As shown in FIG. 2, a head and tail pointer may be maintained for managing the linked list of free buffers.

CPU interface 420 facilitates the configuration of various register and/or memory settings in IRM 410 and/or other modules. Via the CPU interface, IRM 410 is notified of new queue pairs and/or virtual lanes to be established. Illustratively, the IRM is informed of the amount of space needed for the new queue pair or virtual lane (e.g., an initial allocation of buckets) and will reserve that space if it is available, or will assemble or aggregate the necessary space as it becomes available (e.g., as used buckets are returned).

As described in a previous section, NTX 404 transmits outbound communications after they have been reassembled in the shared memory. The NTX notifies IRM 410 when the buckets used by a transmitted communication can be reclaimed.

ILC 406 handles link layer flow control for IRX 402. IRM 410 notifies the ILC of the space available for virtual lanes, and the ILC can then issue an appropriate number of link layer credits on each virtual lane.

More particularly, link level flow control in an embodiment of the invention depends on the allocation of buffers to each virtual lane (for use by the queue pairs that belong to each virtual lane), and the issuance and redemption of credits for traffic traversing the virtual lane. Dynamic programmable threshold registers may be maintained in IRM 410 (i.e., memories 430, 432) or elsewhere, to store the amount of buffer space currently available for each queue pair and/or virtual lane, and or other information.

More specifically, in one embodiment of the invention, queue pair memory (or memories) 430 stores various parameters for managing operation of queue pairs. For example, and as discussed below, dynamic programmable thresholds may be maintained to indicate the amount of buffer space used by (or

available to) a queue pair, programmable amounts of credits a queue pair may advertise depending on which of the thresholds have been reached, whether a queue pair is able to use shared memory buffers, a maximum (if any) amount of dedicated buffer space allocated to a queue pair, etc.

5           Similarly, virtual lane memory (or memories) 432 store operating parameters for virtual lanes, as described below. For example, memory 432 may store the amount of buffer space allocated to each virtual lane.

FIG. 5 demonstrates an apparatus for calculating link level flow control credits, according to one embodiment of the invention. This embodiment is  
10           suitable for implementation with the InfiniBand Resource Manager depicted in FIG. 4. Up to sixteen virtual lanes may be implemented in the illustrated embodiment of the invention.

In FIG. 5, InfiniBand Link Core (ILC) 406 receives, on a per-virtual lane basis, the amount of available buffer space (e.g., in buckets). A virtual lane's  
15           available buffer space may be signaled every time it changes, or with some other regularity. Illustratively, the buffer space available for a virtual lane (or queue pair) may change whenever a new packet or payload is stored (thereby taking up some of the available space) and whenever an outbound communication is transmitted (thereby freeing up space in the shared memory).

20           The FCTBS (Flow Control Total Blocks Sent) value arrives periodically from a link partner, and is noted or stored as "blocks received" 502. After a FCTBS value is loaded, blocks received 502 increments in response to the arrival of blocks from that link partner. A separate "blocks received" value or register may be maintained for each virtual lane.

25           For each virtual lane, ILC 406 adds blocks received 502 and the buffer space available for the virtual lane (e.g., measured in blocks), to calculate FCCL (Flow Control Credit Limit) 504. FCCL 504 represents the maximum amount of

traffic (in credits, one credit per block) the link partner on a particular virtual lane may send on that virtual lane. FCCL 504 is periodically transmitted to the link partner.

5 The link partner, when it wants to send data, determines whether it has any credits available. For example it may subtract the total blocks sent on the link from FCCL. It cannot send the data if the difference is less than or equal to zero.

Because the buffer space that is available for a given virtual lane is continuously updated, and the blocks received on the virtual lane are also continuously updated, FCCL 504 represents the additional amount of traffic that  
10 can be accepted for the virtual lane.

In this embodiment of the invention, for each virtual lane, several values are maintained (e.g., in virtual lane memory 432 of FIG. 4), such as: VL\_enabled, which indicates whether the virtual lane is enabled (e.g., active); VL\_threshold, which identifies the maximum amount of buffer space (e.g., in buckets) that the  
15 virtual lane may use; and VL\_queued, which identifies the current amount of buffer space (e.g., in buckets) that the virtual lane is using. These values are initialized when the virtual lane is established or reset.

As contents of InfiniBand data packets are stored in the shared memory, VL\_queued is incremented and compared to VL\_threshold. If VL\_threshold is  
20 exceeded, the available buffer space for that virtual lane that is reported to ILC 406 by IRM 410 is zero. Otherwise, the difference between VL\_queued and VL\_threshold is reported.

In one alternative embodiment of the invention, the amount of buffer space available for a virtual lane may be calculated by aggregating the buffer space  
25 available for each queue pair within the virtual lane. A method of determining the amount of buffer space available to a queue pair (shared space and/or dedicated space) is described below.



In one embodiment of the invention, the total available buffer space (e.g., in the entire shared memory structure) is divided between virtual lanes based on such factors as the quality of service associated with a virtual lane, arbitration priorities of the virtual lanes, the number of queue pairs belonging to a virtual lane, etc. Not all of the storage space of the shared memory structure need be  
5 allocated to active virtual lanes. Some may be reserved to allow a virtual lane to temporarily exceed its allocation, to facilitate the establishment of a new virtual lane or queue pair, or for some other reason.

Further, the amount of space allotted to a virtual lane is programmable,  
10 and over-subscription of buffer space may be permitted. The amount of buffer space allocated to a management virtual lane (e.g., virtual lane fifteen) may be minimal, such as 384 bytes – enough to store one 256-byte payload plus an InfiniBand header, as required by the InfiniBand specification – or more.

In an embodiment of the invention in which the shared memory structure  
15 is an internal memory 128KB in size, with each bucket being 128 bytes (see FIG. 2), illustrative virtual lane parameter values are as follows. If only two virtual lanes are active – e.g., virtual lanes zero and fifteen – VL\_enabled will be true (i.e., 1) for those two virtual lanes and false (i.e., 0) for all others. For the two active virtual lanes, VL\_queued is initially set to zero. Illustrative VL\_threshold  
20 values are 1,016 (0x3F0) buckets for virtual lane zero and four buckets for virtual lane fifteen. VL\_queued and VL\_threshold are meaningless for inactive virtual lanes.

As another example, if four virtual lanes are active for data (e.g., virtual lanes zero through three), in addition to virtual lane fifteen, illustrative parameters  
25 are as follows. VL\_enabled is true for the active virtual lanes, and false for all others. VL\_queued is zero for virtual lane fifteen and zero for the other active

virtual lanes. VL\_threshold is four buckets for virtual lane fifteen and 252 (0x0FC) buckets for the other active virtual lanes.

In an embodiment of the invention, besides the use of link credits for flow controlling at the link layer, one or more end-to-end flow control mechanisms may be implemented at the transport layer. Illustrative end-to-end flow control mechanisms that may be applied include queue pair credits and RNR-NAK (Retry, Not Ready, Negative Acknowledgement), both of which may be issued on a per-queue pair basis. Allocation of space for RDMA Read responses prior to request may be considered another end-to-end flow control mechanism.

End-to-end credits are issued for each InfiniBand message that the receiver is willing to accept. However, the InfiniBand standard defines a message as being of any length between  $2^0$  and  $2^{32}$  bytes. Therefore, to promote efficient and fair use of the shared memory, the number of dedicated and/or shared memory buckets a queue pair may consume are programmable. Illustratively, as the amount of buffer space (e.g., memory buckets) a queue pair consumes increases past programmable thresholds, the number of end-to-end credits it may advertise is determined by corresponding programmable values. The programmable threshold values and credits may be stored, per-queue pair, in the InfiniBand Resource Manager module (e.g., in queue pair memories 430 of FIG. 4).

Thus, embodiments of the invention provide programmability to allow compromises between disparate goals of good link bandwidth utilization, low retry rate, good buffer utilization, RDMA versus Send modes for data transferred on a queue pair, etc.

FIG. 6 is a graph demonstrating how programmable thresholds and associated queue pair credit values may be arranged, according to one embodiment of the invention. In the graph of FIG. 6, the x-axis represents InfiniBand end-to-end credits (also termed message credits or queue pair credits)

that a node may advertise for a queue pair. The y-axis represents buffer space, which may be measured in buckets.

The graph is divided into two portions, with the lower portion addressing the allocation of dedicated buffer space and the upper portion covering the allocation of shared buffer space. Maximum dedicated threshold 616 separates the two portions, and represents the maximum amount of buffer space dedicated to a particular queue pair. A queue pair may not be authorized or enabled to use shared buffer space and, if not authorized, would therefore not have an upper portion of the graph.

Maximum dedicated threshold 616 may be any value between zero and the size of the shared memory structure, inclusive. Any queue pair having a maximum dedicated threshold of zero uses only shared buffer space, and therefore has no lower portion of the graph and may always contend for buffer space with other queue pairs.

Below maximum dedicated threshold 616, a queue pair is initially able to advertise an amount of credits equal to maximum dedicated credits 610 (e.g., after being established or reinitialized). Once the amount of dedicated buffer space used by the queue pair reaches a first threshold (e.g., dedicated threshold 612a), the number of credits it can advertise decreases (e.g., to dedicated credits 614a). Any number of similar thresholds (e.g., threshold 612b) and credits (e.g., dedicated credits 614b) may be set between zero and maximum dedicated threshold 616.

When the amount of dedicated buffer space used by the queue pair reaches maximum dedicated threshold 616, the queue pair can no longer advertise credits unless it is enabled to use shared buffer space, in which case it advertises credits according to the upper portion of the graph in FIG. 6. Otherwise, it must send RNR-NAK packets directing senders to try sending their data later.

The upper portion of the graph is also marked by a series of thresholds and corresponding amounts of buffer space (e.g., credits). When first allowed access to shared buffer space, a queue pair may advertise an amount of space indicated by maximum shared credits 620. Thereafter, as the amount of used shared buffer space increases past programmed thresholds (e.g., shared thresholds 622a – 622e),  
5 the amount of buffer space or credits the queue pair can advertise decreases accordingly (e.g., shared credits 624a – 624e). In another embodiment of the invention, each queue pair has a “weighting” factor used to modify the advertised shared credit values.

10 The shared threshold values indicated in FIG. 6 measure the total buffer space consumed by all queue pairs authorized or programmed to use shared buffer space. When maximum shared threshold 626 is reached, all queue pairs can no longer advertise credits, and must send RNR-NAKs. In one embodiment of the invention, back-off periods specified in a queue pair’s RNR-NAK packets may be  
15 incremented, in successive packets, until a retry after a particular back-off period succeeds. Upon successful back-off, the back-off period for an individual queue pair may be reset to a default value. The back-off increments may increase linearly, exponentially or in some other fashion.

Note that the amount of credits a queue pair may advertise may increase  
20 when the queue pair exhausts its dedicated buffer space and starts using shared buffer space. That is, maximum shared credits 620 may be greater than a final dedicated credits value (e.g., dedicated shared credits 614b). The number of message credits a queue pair may advertise may be reported to the other end of the queue pair by an acknowledgement generator (e.g., acknowledgement generator  
25 416 of FIG. 4).

Each queue pair’s thresholds and credit limits may differ; therefore, each queue pair’s dedicated region (i.e., lower portion) of the graph may be different.

However, in a present embodiment of the invention, each queue pair enabled to use the shared buffer space applies the same thresholds and credit limits in the shared region (i.e., upper portion).

Various criteria may be considered when selecting and programming appropriate thresholds and credits for a queue pair. And, the memory structure used for buffering queue pair and/or virtual lanes may be under- or over-subscribed. Illustratively, any number of threshold/credit value pairs may be programmed between the extremes of:

$$(Memory\_size / Max\_Packet\_Size) - Max\_Packet\_Size$$

and:

$$(Memory\_size / (Max\_Packet\_Size * Num\_QPs)) - Max\_Packet\_Size$$

where Memory\_size is the size of the shared memory structure (e.g., in buckets), Max\_Packet\_Size is the maximum expected packet size (e.g., in buckets) and Num\_QPs is the number of active queue pairs.

Specific threshold/credit values may depend upon factors such as the type of traffic being received (e.g., encapsulated Ethernet packets), the amount of buffer space dedicated to a queue pair, the amount of shared buffer space, etc. A queue pair's threshold/credit values for its dedicated region of the graph may be stored in a queue pair-specific memory (e.g., in the IRM module). The shared threshold/credit values for the shared region may be stored in a globally accessible register or structure.

In one embodiment of the invention, the maximum dedicated credits value for all queue pairs that are active and not dedicated to management traffic may be set to the maximum amount of buffer space – e.g., 1,024 buckets in the shared memory structure of FIG. 2. A first dedicated threshold may be set at one bucket, to decrease the number of advertisable credits to 768. A second dedicated threshold may be set at two buckets, to decrease the number of advertisable

credits to 512. The active, non-management queue pairs' maximum dedicated thresholds may be set to three buckets, and all may be enabled to use shared buffers.

5 In this embodiment, the maximum shared credits value may be 1,023, while the maximum shared threshold value may be 831, to leave shared buffer space for completing the reassembly of outbound communications, facilitating establishment of a new queue pair or virtual lane, or for some other purpose. Illustrative shared threshold/advertisable credit tuples may be as follows: <64, 192>, <128, 96>, <256, 48>, <576, 24>, <704, 3> and <768, 1>.

10 In the embodiment of the invention illustrated in FIG. 6, RNR-NAK is a primary method of flow controlling queue pairs. An RNR-NAK will be sent if a queue pair's dedicated buffer space is full and it is not permitted to use shared buffers, or if the queue pair's dedicated buffer space is full, and it is permitted to use shared buffers, but the shared buffer space is full.

15 To facilitate an RDMA Read operation, resources may be reserved for the operation before the RDMA Read is issued. But first the IRM must determine whether the corresponding queue pair has sufficient space. Illustratively, the queue pair's depth is calculated as the number of buckets currently in use (i.e., to store payloads of received InfiniBand packets) plus the number of buckets  
20 currently reserved for outstanding RDMA Reads. A new RDMA Read will only be issued if this sum and the queue pair's programmable thresholds indicate that buffer space is available. Space for the RDMA Read will then be allocated from the dedicated and/or shared buffer space.

25 In addition to the programmable threshold and credit values, the number of outstanding RDMA Read operations a queue pair may have may also be programmable, and may be maintained in the IRM or some other module. Other parameters the IRM may store for a queue pair include the queue pair's current

depth, whether the last packet received for the queue pair resulted in an RNR-NAK, whether increasing back-off is enabled for RNR-NAK, etc.

In one embodiment of the invention, there is a maximum number of outbound communications that may be reassembled in the shared memory structure at one time. When this maximum number is reached, no new queue pairs (or virtual lanes, in one implementation) may be established until the number of communications being reassembled falls below the maximum. Thus, in this embodiment, a programmable amount of buffer space is reserved for completing the reassembly of communications, and the maximum shared threshold value (e.g., threshold 626 in FIG. 6) will be less than the total amount of shared buffer space in the memory structure.

Illustratively, when used buffer space (e.g., one or more memory buckets) is returned, the returned space is first added to the amount of shared space that is available. Then, when the shared space has been fully restored, or restored past a specified threshold (e.g., dedicated threshold 616), space may be returned to individual queue pairs' dedicated regions. The manner in which returned buckets are added to shared or dedicated buffer space may differ in other embodiments of the invention.

FIG. 7 demonstrates a method of performing link layer and transport layer flow control for InfiniBand traffic received at a channel adapter, according to one embodiment of the invention. The various virtual lanes and queue pairs share a single memory structure comprising multiple buckets or buffers that can be dedicated to a virtual lane or queue pair, or that can be part of a pool of shared buckets.

In operation 702, various operating parameters are set that pertain to the virtual lanes and queue pairs terminating at the adapter. For a virtual lane, the parameters may include the amount of buffer space currently used, the maximum

amount that may be used (e.g., the number of buffers dedicated to the virtual lane), and whether the virtual lane is active.

For a queue pair, the parameters may include indications of whether the queue pair is active and whether it can use the pool of shared buckets. Other  
5 possible queue pair parameters include the various dedicated and shared thresholds and matching credit values, the maximum number of credits that may be advertised when using dedicated and/or shared buckets, an RNR-NAK timer and back-off enable value, a measure of the current amount of memory space (e.g., number of buckets) used by the queue pair, etc. The various parameters may  
10 be stored in an InfiniBand resource manager module.

In operation 704, a new packet is received, and its virtual lane and queue pair are identified (e.g., by the resource manager or a receive packet processor module). Some or all of the operating parameters associated with the virtual lane and/or queue pair may be retrieved.

15 In operation 706, the size of the packet payload is determined. In this embodiment, only the packet payload is to be stored in the shared memory structure. In other embodiments, other contents – such as the entire packet – may be stored, in which case the size of the whole packet is noted.

In operation 708, it is determined whether the virtual lane's threshold (e.g.,  
20 maximum size) would be exceeded if the payload were stored. If not, the method advances to operation 712; otherwise, the method continues with operation 710.

In operation 710, the packet is dropped or rejected. The method then ends.

In operation 712, the queue pair is examined to determine if it can accept the payload. For example, the current size of the queue pair's queue (e.g., the  
25 number of buckets it uses) may be compared to its maximum number of dedicated buckets. Further, if the queue pair is permitted to draw from the pool of shared buckets, it may be determined whether any shared buckets are available. If the



queue pair can store the payload, the illustrated method advances to operation 718.

5 In operation 714, it is determined whether the queue pair is a reliable connected queue pair (RCQP). Alternatively, the queue pair may be an unreliable datagram queue pair (UDQP). If the queue pair is an RCQP, the method continues with operation 716; otherwise, the method concludes with operation 710.

In operation 716, an RNR-NAK response is sent to the other end of the queue pair, to direct the sender to retry the packet later. The method then ends.

10 In operation 718, the packet payload is stored and the virtual lane and/or queue pair data are updated accordingly. For example, their queue sizes may be incremented, the linked list for the queue pair may be adjusted (e.g., if a new bucket was put into use), etc.

15 In operation 720, the number of link and transport layer credits remaining for the virtual lane and queue pair are calculated. A method of determining available link layer credits is described above in conjunction with FIG. 5. The number of message credits that can be advertised for the queue pair may be determined by identifying the lowest dedicated or shared threshold the queue pair's queue size has exceeded, and retrieving from storage the number of credits 20 associated with that threshold.

In operation 722, the link layer and transport layer credits are advertised to the link partner and connection partner, respectively. The method then ends.

## 25 **Mapping Communications to an External Port, with Combined Buffering of Virtual Lanes and Queue Pairs**

In an embodiment of the invention, a system and methods are provided for mapping communications between an InfiniBand queue pair (QP) and a

communication system or component external to the InfiniBand fabric. For example, a fabric may be coupled to an external network or communication system (e.g., an Ethernet network) via a computing or communication device comprising an InfiniBand channel adapter, as shown in FIG. 1.

5           Within the device, communications destined for the external system (e.g., Ethernet packets) are received from the InfiniBand fabric via various queue pairs and virtual lanes, stored in a shared memory structure, and then transferred to the external system. FIG. 2 depicts one implementation of the shared memory structure.

10           In this embodiment, the shared memory structure is used to reassemble traffic as it is received from the fabric. The memory structure is shared among multiple queue pairs, each of which may have traversed any virtual lane (VL) to arrive at the computing device. As described above, the shared memory structure is managed as a set of linked lists, thereby allowing traffic from multiple queue  
15           pairs and virtual lanes to use the same structure simultaneously. When a complete communication is reassembled, the portion of the corresponding queue pair's linked list that encompasses the communication is transferred to a module configured to process the communication for transmission (e.g., NTX 130 of FIG. 1). This avoids the need to copy the communication between the inbound  
20           InfiniBand queue pair/virtual lane and the outbound port or connection.

          In the embodiment of the invention illustrated in FIG. 2, when an outbound communication is fully reassembled in memory 204, pointers identifying the buckets in which the communication is stored are copied to transmit pointers 220 for the appropriate external port. Each external port may  
25           have one or more associated queues (e.g., for different qualities of service).

          For example, if the reassembled communication was received on queue pair 2 and is intended for transmission on port A, the portion of queue pair 2's

linked list encompassing the communication is moved to the linked list anchored by PortA head 222a and PortA tail 222b. Thus, transmit pointers 220 enable access to control 202 and memory 204 for the transmission of communications, and the control and memory serve not only as an input queue from the InfiniBand fabric, but also an output queue for external communications.

The sizes of the queue pair and external port linked lists are dynamic, allowing them to grow and shrink as InfiniBand traffic is received and as outbound communications are transmitted. The outbound communications may be received via InfiniBand Send commands, RDMA Read operations and/or other forms.

In an embodiment of the invention, a new outbound communication is received encapsulated within one or more InfiniBand packets. When reassembly of the new communication is initiated, the first memory bucket used for the new communication is also used to store various meta-information before storing the first portion of the communication. Illustratively, the meta-information may be stored in the first line of the first bucket, and may include information such as: the virtual lane and queue pair through which the communication is received, the external port through which it should be transmitted, a quality of service (QoS) to be afforded the communication, a checksum calculated on the communication, checksum offset, etc.

When control of the reassembled communication is passed to a network transmit module, the port and QoS information indicate where the pointers to the linked list describing the communication should be queued. The virtual lane and queue pair information may be used by the network transmit module, after the communication is transmitted, to return the used memory buckets (e.g., to the InfiniBand resource manager or receive module).

To properly manage the queue pairs and the issuance of queue pair (and/or virtual lane) credits, buckets in which an outbound communication is reassembled may not be released from a queue pair's linked list until the communication is transmitted and the external port releases the associated control entries and  
5 buckets. Then one or more parameters (e.g., depth counters) indicating the amount of the shared memory structure used by the queue pair (and/or virtual lane) are updated, which may allow the queue pair (and/or virtual lane) to advertise more credits.

The queue pairs through which communications are received for  
10 reassembly may be reliable connected queue pairs (RCQPs). Thus, each InfiniBand packet received in one of the RCQPs may have a packet sequence number, and ordered receipt of the fragments of an outbound communication can be ensured.

Some or all of the components illustrated in FIG. 2, as well as other  
15 elements of a channel adapter, may be embedded on a single ASIC (Application-Specific Integrated Circuit). Thus, the size of memory 204 and control 202 may be constrained. The use of linked lists for the queue pairs allows flexible and efficient use of buffering memory, without wasting space or denying memory buckets to a queue pair or virtual lane when needed.

20 FIG. 8 demonstrates a method of mapping communications between a memory structure shared among multiple inbound InfiniBand queue pairs and one or more outbound communication ports, according to one embodiment of the invention.

In operation 802, an InfiniBand packet is received at a channel adapter of a  
25 device coupled to an InfiniBand fabric and an external communication system (e.g., an Ethernet network), and is validated according to the InfiniBand specification. This embodiment may be particularly suited for processing

InfiniBand Send commands that comprise encapsulated communications (e.g., Ethernet packets). Alternatively, the packet may contain a response to an RDMA Read operation or some other content to be reassembled into an outbound communication.

5           In operation 804, the virtual lane and queue pair on which the InfiniBand packet was received are identified. Illustratively, the service level associated with the packet's virtual lane may affect the QoS afforded the Ethernet packet when it is transmitted on one of the outbound ports.

          In operation 806, an InfiniBand resource manager or other module  
10       determines whether there is space for the packet's payload in the shared memory structure. In particular, the resource manager may determine whether storing the payload would exceed the memory (e.g., buckets) allocated or available to the packet's virtual lane or queue pair.

          As described in a previous section, various parameters may be maintained  
15       to reflect the amount of storage space used by a virtual lane or queue pair, the amount of space dedicated to a virtual lane or queue pair, the amount of shared space available, whether a virtual lane or queue pair is permitted to use the shared space, etc. If the payload cannot be accommodated, the packet is dropped and/or an RNR-NAK notification is sent to the other end of the connection, and the  
20       illustrated method ends.

          In operation 808, the payload of the InfiniBand packet is stored in the shared memory structure. To find the location where the payload should be stored, the linked list of memory buckets associated with the packet's queue pair is referred to. For example, a tail pointer can be followed, which points to the last  
25       bucket currently allocated to the queue pair.

          In this embodiment, entries in the shared control structure for the shared memory structure contain information regarding the status of the bucket that

corresponds to the entry. Thus, it can readily be determined where in a bucket the payload should be stored and whether the entire payload will fit in the bucket.

Illustratively, if the payload comprises the first fragment or portion of an outbound communication, a new bucket may be used. Otherwise, the payload is stored contiguously with the previous payload for the communication.

More specifically, if a newly arrived InfiniBand packet is part of a partially reassembled outbound communication, it is used to fill the remainder of the tail bucket (i.e., the bucket pointed to by the linked list's tail pointer). If the tail bucket is full, or the InfiniBand packet is part of a new message, then a new bucket is started. In this case, the new bucket is pointed to by the address in the control structure pointed to by the tail pointer, and the new bucket becomes the new tail bucket.

As described above, the first bucket for a new communication may be prepared by first storing information that will facilitate mapping of the communication to an external port and reclamation of the used memory buckets.

Also in operation 808, the queue pair's linked list is updated as necessary. For example, if a new memory bucket is needed to accommodate the payload, the bucket and a new corresponding entry in the control structure are initialized as necessary and added to the queue pair's linked list.

In operation 810, depth indicators (e.g., pointers, counters) are updated as necessary to indicate how much traffic is currently stored in the shared memory for the queue pair and/or the associated virtual lane. Illustratively, these parameters are maintained by the resource manager.

In operation 812, it is determined whether a full outbound communication (e.g., Ethernet packet, Packet Over SONET/SDH) has been reassembled. If so, the illustrated method advances to operation 814; otherwise, the method returns to

operation 802 to accept another InfiniBand packet, for the same or a different queue pair.

5 In operation 814, the sub-list (of the queue pair's linked list) is identified and posted or copied to the linked list structure corresponding to the outbound port through which the completed communication will be transmitted. The sub-list comprises a set of linked memory buckets and corresponding control entries.

10 To facilitate this re-mapping of the sub-list, pointers to the first and last elements of the sub-list may be passed from the InfiniBand receive module (IRX) to the network transmit module (NTX) after the last portion of the communication is reassembled. These pointers will identify the first control entry and memory bucket and the last control entry and memory bucket. In the illustrated method, the sub-list is not yet removed from the queue pair's linked list.

In operation 816, the outbound port or the NTX links the sub-list to a linked list for the appropriate outbound queue of the external port.

15 In operation 818, the outbound port transmits the communication. Various processing may be performed before the transmission (e.g., checksum calculation, adding headers, VLAN ID insertion) by the NTX and MAC (Media Access Control) module or other component(s).

20 In operation 820, the outbound port (NTX) releases the control entries and memory buckets corresponding to the communication (e.g., back to the IRM). Illustratively, control entries may include flags or other indicators revealing whether a control entry (and/or associated memory bucket) is currently used by an outbound port. The flags will be reset when the associated memory buckets are reclaimed.

25 In operation 822, the control entries and memory buckets are released from the queue pair and returned to the free list. The memory buckets and/or control entries may be flushed to clear their contents.

In operation 824, depth indicators for the queue pair (and/or virtual lane) through which the communication was received are updated to reflect the release of the sub-list.

5 In operation 826, queue pair and/or virtual lane credits may be issued, if possible, depending on how much traffic in the shared memory structure now corresponds to the queue pair and virtual lane (e.g., the queue pair or virtual lane depth and the dedicated or shared thresholds). The issuance of credits may also depend upon how many free buckets are available, how many buckets may be reserved for other queue pairs, how many credits have been issued by other queue  
10 pairs and/or virtual lanes, etc.

For example, and as described in a previous section, a queue pair may or may not be allocated a number of dedicated memory buckets, and may or may not be permitted to use a pool of shared buckets. Thresholds may be defined within the dedicated and/or shared sets of buckets. Different numbers of message credits  
15 may be advertisable as a number of buckets used by a queue pair meets or exceeds each threshold level of buckets. Thus, to determine if (and how many) message credits the queue pair can now advertise, the resource manager (or other module) may compare the present number of buckets used by the queue pair with applicable thresholds.

20

#### **Interleaving Sends and RDMA Read Responses in a Single Receive Queue**

In one embodiment of the invention, an InfiniBand queue pair's receive queue receives and interleaves different types of traffic at a channel adapter. For example, the queue may receive Send commands that convey encapsulated  
25 outbound communications (e.g., Ethernet packets). Each Send command may comprise any number of individual InfiniBand packets, the payloads of which are re-assembled in the queue to form the communication.



The queue may also receive Send commands that convey one or more RDMA Read descriptors. When processed, the descriptors are used to configure RDMA Read requests from the channel adapter to the host that originated the Send command. In response, the host sends RDMA Read responses conveying portions of an outbound communication to be assembled and transmitted by the channel adapter.

Illustratively, the queue pair queue may be implemented as a series of linked buckets in a shared memory structure, as described in preceding sections. In other embodiments, the queue may be implemented in other ways, and may be configured to store traffic other than InfiniBand Sends and Reads, such as PCI Express communications, Asynchronous Transfer Mode (ATM) traffic, Fibre Channel transactions, etc.

FIG. 9 depicts a queue pair queue as it may be configured to store and facilitate the processing of mixed InfiniBand receive traffic, in one embodiment of the invention. In this embodiment, queue pair queue 902 may be conceptually viewed as comprising a single queue, and may be treated as a single queue for flow control purposes. However, in this embodiment of the invention, queue 902 comprises multiple linked lists, each with a head and tail pointer.

Thus, queuing area 910 is where new InfiniBand Send command packets are received and queued for processing (e.g., for reassembly or forwarding of the outbound communication corresponding to the Send command). Queuing area 910 has associated head pointer 912 and tail pointer 914. Tail pointer 914 identifies a tail of the queue, which may indicate where the next new content (e.g., new Send command) is to be stored.

Assembly (or reassembly) area 920 is where outbound communications are assembled (or reassembled) from portions received via separate InfiniBand packets (e.g., Send commands encapsulating an outbound communication,

responses to RDMA Read requests). The assembly area is managed with the aid of assembly head pointer 922 and assembly tail pointer 924. The head of the assembly area of queue 902 coincides with the first bucket of the next communication to be passed to a Network Transmit (NTX) module.

5           Queuing area 910 and/or assembly area 920 may be expanded by drawing upon a free list of buffers; as buffers are used in queue 902, they are returned to the free list (e.g., after their corresponding outbound communications have been transmitted).

10           In the illustrated embodiment of the invention, head pointer 912 identifies the entry or element (e.g., the next Send command) in queue 902 that corresponds to the next outbound communication to be assembled, reassembled or passed forward.

15           In this embodiment of the invention, new Send commands (encapsulating an outbound communication or a set of RDMA Read descriptors) are enqueued at tail pointer 914 of the queuing area of queue pair queue 902. As Send command packets are processed or dispatched from the queuing area to the assembly area, head pointer 912 is adjusted accordingly. As reassembled communications are forwarded to the NTX module, assembly head 922 and assembly tail 924 are advanced as needed.

20           When head pointer 912 is advanced to a new command in queuing area 910, the following action depends on which type of traffic the command is. If it is a Send command encapsulating an outbound communication, the communication should be fully re-assembled by the time the head pointer gets to the command. If so, then the communication can simply be passed to the NTX module. This may  
25           be done by forwarding pointers identifying the beginning and end of the communication, as described in a previous section. The buckets containing the

communication are then unlinked from queue pair queue 902 and added to the external port's queue.

If the command identified by the next packet pointer is a Send command encapsulating a set of RDMA Read descriptors, the Read descriptors are retrieved  
5 and corresponding RDMA Read requests are issued. As they are received, the resulting RDMA Read responses bypass the tail of the queue and are stored directly in assembly area 920. Buckets may be added to the tail end of the assembly area of the queue as necessary to accommodate the responses.

Thus, queuing area 910 and assembly area 920 may be seen as separate  
10 queues or sub-queues. However, the last bucket of assembly area 920 may reference head pointer 912 or the first bucket of queuing area 910.

FIG. 10 demonstrates in greater detail a method of handling mixed InfiniBand receive traffic (e.g., Sends and RDMA Read responses) in a single queue pair receive queue.

15 In operation 1002, InfiniBand Send packets conveying various content are stored in the queue. Head, tail and/or other pointers of the queuing area of the queue (e.g., a next packet pointer) are updated accordingly. Illustratively, most or all information other than the payloads is stripped from the InfiniBand packets before they are enqueued. The information that is stripped off may include one or  
20 more headers, CRCs, etc.

In one embodiment of the invention, each section of the queue pair queue has a first and last pointer configured to identify the beginning (e.g., first memory bucket) and end (e.g., last bucket) of the section. These pointers may facilitate the transfer of communications and other content, and may overlap or be separate  
25 from the head, tail and next packet pointers discussed above.

In operation 1004, the next packet pointer is advanced to a queue entry containing an unprocessed command or packet. The queue entry may comprise one or more memory buckets.

5 In operation 1006, the type of command or traffic is identified. This may be done by reading a sufficient portion of the command or packet. For example, the payload of the first InfiniBand packet conveying a new command may include meta-data, a proprietary header or other information indicating what the command is.

10 If the command is an encapsulated Send (i.e., a Send command conveying an encapsulated outbound communication), the method advances to operation 1012. If the command is a Send command conveying one or more RDMA Read descriptors, the illustrated method continues at operation 1008.

15 In operation 1008, the RDMA Read descriptors are identified and corresponding RDMA Read requests are issued to the specified host (e.g., one Read request per descriptor) to retrieve the content of a communication. In addition, one or more buckets may be appended to the assembly area of the queue pair queue, if needed (e.g., based on the expected size of the communication).

20 In operation 1010, RDMA Read responses, corresponding to the RDMA Read requests and containing portions of the communication, are received and assembled in the assembly area of the queue. The communication portions may be shifted or adjusted as necessary to store them contiguously. This may also require one or more queue pointers to be updated.

25 In operation 1012, it is determined whether all RDMA Read responses have been received for the current communication. If so, the method continues with operation 1014; otherwise, the method returns to 1010 to continue receiving the responses.

In operation 1014, the RDMA descriptors are returned.

In operation 1016, it is determined whether the communication is complete (i.e., completely assembled or re-assembled). If not, the illustrated method returns to operation 1002. While the communication is being assembled or reassembled, a checksum may be calculated. The checksum, checksum offset  
5 and/or other information useful for further processing of the communication (e.g., external port number, quality of service) may be stored in front of the communication (e.g., the first lines of the first bucket) for use by the NTX module.

In operation 1018, the completed communication is dispatched to the NTX  
10 module. Illustratively, the first and last pointers of the assembly area are used to unlink the completed communication's buckets from the queue pair queue and link them to the corresponding external port or a particular queue of the external port. Queue pointers are updated as needed, and the illustrated procedure then ends.

15

### **Tracking Responses to an RDMA Read Operation**

In one embodiment of the invention, responses to an RDMA Read operation are tracked in the channel adapter that issued the operation. Illustratively, for each queue pair or other communication connection in which  
20 RDMA Reads are performed, a linked list or other structure (e.g. a FIFO) is maintained. A single memory structure may be employed, in which the linked lists may be interleaved.

In this embodiment, each entry in a queue pair's linked list identifies the range of Packet Sequence Numbers (PSN) associated with the expected responses  
25 to one RDMA Read operation. Each entry also includes a link to the next entry in the list, and may also store the PSN of the most recently received response. When the last response is received, the entry may be removed, and a retry queue entry

corresponding to the RDMA Read may be retired. However, if a response is received out of order, the retry queue entry may be used to retry the RDMA Read.

FIG. 11 depicts a memory structure for maintaining the linked lists in a channel adapter, according to one embodiment of the invention. In this embodiment, RDMA PSN tracking memory 1102 includes any number of entries. The size of memory 1102, and number of entries, may depend upon factors such as the number of queue pairs established on the channel adapter, the number of queue pairs enabled to perform RDMA Read operations, the average or expected number of RDMA Reads, the estimated period of time needed to complete an RDMA Read, etc.

RDMA PSN tracking memory 1102 may include virtually any number of linked lists, but can accommodate one linked list for every active queue pair that performs an RDMA Read operation.

Memory 1102 of FIG. 11 includes one linked list, which starts with entry m and also includes entries n and 1. Each entry corresponds to a single RDMA Read operation. Each entry may be configured as shown for entry m, which identifies the first and last PSNs of the expected responses to the RDMA Read, the PSN of most recently received (i.e., latest) response, and a link to the next entry in the linked list.

Any entry in memory 1102 may be part of any linked list. However, a particular entry (e.g., entry 0) may be a null entry and may be used to terminate a linked list.

For each linked list in RDMA PSN tracking memory 1102, pointers are maintained to the first and last entries. These pointers may be stored in memory or registers, such as RDMA tracking registers 1110, or some other structure. Thus, in FIG. 11, QPi head pointer 1112a identifies the head of the linked list for queue pair i, while QPi tail pointer 1112b identifies the tail of the linked list.

Free entries – entries in memory 1102 not being used for any linked lists – may be organized into a free list. In FIG. 11, a free list is maintained using free head 1104a, which identifies a first free entry, and free tail 1104b, which identifies a last free entry. The free entries may be linked using the “link” field of the entries; the other fields are not used.

In other embodiments of the invention, an entry in a queue pair’s linked list may contain virtually any information.

FIG. 12 demonstrates one method of tracking responses to an RDMA Read operation using a linked list, according to one embodiment of the invention.

In operation 1202, a channel adapter (for an InfiniBand environment) or other communication interface (for some other environment) receives one or more RDMA Read descriptors. For example, a channel adapter may receive, on a particular queue pair, a Send command containing one or more RDMA Read descriptors. Each descriptor describes all or a portion of a communication (e.g., an Ethernet packet) to be retrieved from a host. In this embodiment, the queue pair is a reliable connected queue pair.

In operation 1204, a transmit module of the channel adapter eventually issues an RDMA Read operation to retrieve the communication, or communication portion, from the host. The amount of data to be retrieved may be of virtually any size. The RDMA Read operation is assigned a Packet Sequence Number (PSN) to distinguish it from other transactions conducted by the channel adapter.

In operation 1206, the number of expected responses to the RDMA Read is calculated. This may be done by dividing the amount of data to be retrieved by the MTU (Maximum Transfer Unit) size in effect for the queue pair (or other communication connection) through which the data will be retrieved. As one skilled in the art will appreciate, each response will have a different PSN, and all

the responses' PSNs should be in order, from a first PSN for a first response to a last PSN for a last response. Thus, the PSNs of the expected responses can be readily ascertained. This range of PSN values is communicated to a receive module that will receive the responses.

5           In operation 1208, the receive module initializes or adds to a linked list (such as the linked list of FIG. 11) corresponding to the queue pair. The receive module populates a new entry in the queue pair's linked list with the range of PSNs (e.g., first and last) and a link to the next (or a null) entry in the linked list. A "latest" field in the entry, for tracking the PSN of the most recently response,  
10       may be initialized to a null or other value.

          In operation 1210, an RDMA Read response is received on the same queue pair on which the RDMA Read was issued in operation 1204. As explained in a previous section, RDMA Read responses may be received interleaved in a receive queue with other traffic (e.g., InfiniBand Sends).

15           In operation 1212, the PSN of the RDMA Read response is identified (e.g., by reading it from a header of the response).

          In operation 1214, the PSN is examined to determine if it is out of order. If the response is for the issued RDMA Read, but is not the next expected response, the method advances to operation 1218.

20           In operation 1216, the PSN is in order, and a determination is made as to whether this response is the last one expected for the RDMA Read. If not, the method returns to operation 1210 to receive the next response.

          Otherwise, in operation 1218, the transmit module is notified that all responses have been correctly received. Therefore, the transmit module, if it  
25       maintains a retry queue for RDMA Reads, can retire a retry queue entry corresponding to the RDMA Read issued in operation 1204.



However, if operation 1218 is reached directly from operation 1214, the transmit module is notified of the out of order receipt of a response to the RDMA Read. The transmit module may then retry the operation. In this case, the corresponding entry in the linked list may be removed in favor of the retried operation.

The foregoing embodiments of the invention have been presented for purposes of illustration and description only. They are not intended to be exhaustive or to limit the invention to the forms disclosed. Accordingly, the scope of the invention is defined by the appended claims, not the preceding disclosure.